university of groningen

University Medical Center Groningen

University of Groningen

The role of working memory in young second language learners' written performances

Michel, Marije; Kormos, Judit; Brunfaut, Tineke; Ratajczak, Michael

Published in:
Journal of Second Language Writing

DOI:
10.1016/j.jslw.2019.03.002

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2019

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Michel, M., Kormos, J., Brunfaut, T., & Ratajczak, M. (2019). The role of working memory in young second language learners' written performances. *Journal of Second Language Writing*, *45*, 31-45. https://doi.org/10.1016/j.jslw.2019.03.002

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 05-05-2023

Journal of Second Language Writing 45 (2019) 31–45

Contents lists available at ScienceDirect

Journal of Second Language Writing

journal homepage: www.elsevier.com/locate/jslw

The role of working memory in young second language learners' written performances

Marije Michel[a,b,*], Judit Kormos[a], Tineke Brunfaut[a], Michael Ratajczak[a]

[a] Lancaster University, UK
[b] Groningen University, the Netherlands

ARTICLE INFO

Keywords:
Working memory
Young learners
Writing assessment
Testing writing
Individual differences

ABSTRACT

This study investigated the role of working memory (WM) in the second language (L2) writing performance of young English language learners. It also examined how L2 writing achievement relates to task type and grade level and whether the effect of cognitive abilities varies across different task types and grade level. The participants were 94 young learners (Grades 6 and 7) in Hungary, who performed four writing task types as part of the TOEFL® Junior™ Comprehensive test-battery and completed cognitive tests that assessed their WM functions. Participants scored high on the email writing and integrated Listen-Write tasks. Irrespective of WM functions, on average learners in Grade 7 outperformed those in Grade 6 on the Listen-Write task and the Email task. Students gained lower scores on the non-academic version of an editing task than on most other types of tasks. WM functions had no significant relationship with L2 writing scores, except for the academic editing task. In Grade 7, the effect of WM was not significant on the integrated Listen-Write task, but it resulted in the change of expected score. Learners with high working memory in Grade 6 showed somewhat more consistent performance across tasks than did learners with low working memory.

1. Introduction

A large number of second language (L2) learners in instructed settings are young learners (Butler, 2017), as foreign languages are often a compulsory school subject. In particular, English, being a lingua franca, is regarded as an important target language because of its key role in international education and employment opportunities. Given the vital role of English for the future of young learners, the teaching and assessment of English language skills for this age group has recently become the focus of growing interest (Nikolov, 2016; Wolf & Butler, 2017). In the construction and evaluation of teaching tasks and assessment tools for young learners, special attention is being paid to the developing cognitive capacity and affective characteristics of younger learners (Bailey, 2017). With a focus on assessment, research has also begun to investigate the age appropriateness, validity and reliability of young learners' tests in various educational contexts (Papageorgiou & Cho, 2014; Papp & Walczak, 2016; Wolf & Butler, 2017).

The present study contributes to this line of research by exploring the role of working memory (WM) functions in the writing performances of young L2 English learners, adopting a cognitive perspective on L2 writing (cf. Cumming, 2016). Despite a growing body of research into the relationship between WM functions and adult L2 performance (see meta-analysis by Linck et al., 2014), we still have a limited understanding of how young learners' WM relates to their L2 task performance. Yet, individual differences in WM

* Corresponding author at: Groningen University, Department of Applied Linguistics, Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, the Netherlands.
E-mail address: m.c.michel@rug.nl (M. Michel).

https://doi.org/10.1016/j.jslw.2019.03.002
Received 22 May 2018; Received in revised form 3 March 2019; Accepted 3 March 2019
1060-3743/ © 2019 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/BY/4.0/).

are particularly relevant for children whose attentional regulation mechanisms are still developing (Jarvis & Gathercole, 2003). Study of the link between WM storage capacity and attention regulation ability is also important because teaching and assessment tasks can vary in their demands on attentional resources (Robinson, 2001). Hence, tasks that are excessively taxing on the storage and processing functions of WM and attentional resources might not contribute to children's L2 development and might unfairly disadvantage children with lower levels of WM functioning in high-stakes and classroom testing contexts. Therefore, it seems critical to examine whether certain task types, especially more complex ones which involve the integration of writing with skills such as reading or listening, create an unnecessarily high working memory load and thereby restrict the potential of L2 learning through writing (Byrnes & Manchón, 2014). As assessment of L2 writing skills is also a critical part of writing pedagogy that can provide diagnostic information, it is also important to examine how working memory limitations can result in construct-irrelevant variance in the assessment of young L2 learners.

In the present study, 94 L2 English learners in Hungary, between 11 and 14 years of age, completed the writing section of the computer-administered TOEFL® Junior™ Comprehensive test battery. A key objective of this test is to equip teachers with valid and reliable information about young learners' L2 skills that can inform future instructional design. The writing component of the test comprises four task types: two editing tasks, one email writing task, an opinion essay and a listen-write task in which students have to write a summary of an aural text. In addition, the participants of this study were administered three age-appropriate WM tasks that assessed their phonological short-term memory, the storage, and the processing and central executive (CE) functions of WM. Cumulative Link Mixed Models (CLMMs; Agresti, 2010) were used to examine the relationships between the learners' writing test scores and WM functions and potential changes in this relationship depending on task type and grade level.

In the following, we will first review research on the role of WM functions in L2 writing with a focus on young learners. Next, we will describe the research design of our study and data collection procedures. We then report the results of the CLMMs and discuss them in light of theoretical perspectives on the English-L2 writing performance of young learners. We conclude with practical implications of our findings for the teaching and evaluation of L2 writing for young learners.

2. The role of WM in L2 writing

2.1. Writing and WM models

Writing is a complex, meaning-making cognitive process in which a variety of social and cognitive factors play a role (Byrnes & Manchón, 2014, MacArthur & Graham, 2016). From existing cognitive accounts describing writing processes (e.g., Hayes & Flower, 1980; Bereiter and Scardamalia, 1987; see review by Cumming, 2016), we adopted Kellogg's (1996) writing model in our study since this model has been successfully employed in prior L2 research (e.g., Kormos, 2012; Manchón, Murphy & Roca de Larios, 2009; Révész, Michel, & Lee, 2017) and explicitly links writing processes to WM functions (see Olive, 2012, for a review on L1 writing). Kellogg (1996) distinguishes three highly interactive and recursive sub-processes: formulation, execution and monitoring. During formulation, writers plan the content of their text by retrieving ideas from long-term memory and translate this content into linguistic form by drawing on the processes of lexical retrieval, syntactic encoding and the expression of cohesive relationships. The execution stage involves actual motor movements (holding and moving a pen, typing on a keyboard). Monitoring entails revision and editing behaviour to ensure that the composed text complies with the intentions of the writer. Among others, Bourdin and Fayol (2002) showed that each of these sub-processes draws heavily on attentional resources.

WM is responsible for the short-term storage, active processing, and manipulation of information in the cognitive system, and it is therefore a highly relevant construct for writing processes (see reviews by Olive, 2012 and MacArthur & Graham, 2016). In our research, we defined WM according to Baddeley's (2003) highly influential model (adapted from Baddeley & Hitch, 1974). This specifies WM as consisting of a Central Executive and three domain-specific slave-systems for phonological, visual/spatial, and episodic information, respectively. The phonological loop is responsible for the short-term retention and manipulation of verbal information, the visuo-spatial sketchpad stores and processes visual and spatial information, while the episodic buffer merges smaller pieces of information into episodes. The CE controls attentional processes, such as focusing, dividing and switching attention. It is also responsible for the activation and inhibition of processing routines and regulates how information is exchanged between short-term slave-systems and the long-term memory system. An important assumption of this WM model is that the CE as well as the three slave-systems are limited in capacity. Given that all information will first be processed by WM before it may enter long-term memory, its limited capacity "acts as a bottleneck for learning" (Gathercole & Alloway, 2008, p.12).

Earlier research into the role of WM in L2 processing and learning has demonstrated that individuals with high WM capacity achieve higher language proficiency, have a larger vocabulary size, and are in general more competent users of the four language skills (e.g., meta-analysis: Linck et al., 2014; reviews: Juffs & Harrington, 2011; Kormos, 2012). In addition, individual differences in the storage and processing functions of WM and attention regulation have been shown to play a role in how learners perform on language tests. For example, Mitchell, Jarvis, O'Malley, and Konstantinova (2015) found a strong relationship between advanced learners' TOEFL iBT scores and WM capacity.

2.2. Research on the role of WM in L2 writing

Given that writing involves complex interactive and recursive cognitive processes, Kormos (2012) argued that individual differences in WM functioning are likely to influence how L2 learners manipulate and store information during text production. Even though writing is less time-constrained than speaking, it still involves the simultaneous retrieval, storage and manipulation of

language and ideas to be conveyed. In developing writers, such as young learners, the mechanical processes of writing letters – be it by hand or on a key-board – are also likely to need some attention and mental effort (Berninger, 1999; Olive, 2012). In sum, it is expected that limitations in the storage, processing and attention regulation functions of WM might impact all stages of the writing process (Kormos, 2012). Johnson (2017) hypothesized that larger phonological short-term memory (PSTM) storage capacity might allow for the processing of longer and more complex lexical units and syntactic structures. He also assumed that visuo-spatial short-term memory assists in planning and monitoring processes, since these are related to the processing of visual information (Kellogg, 1996). Individuals with larger WM storage capacity and more efficient CE functioning are also expected to benefit during all stages of writing because the coordination of parallel processing of information and switching between sub-tasks of text composition all draw heavily on the availability of attentional resources (Révész et al., 2017).

While several studies have explored the role of WM functions for L1 writing in adults (e.g., Hoskyn & Swanson, 2003; Olive, Kellogg, & Piolat, 2008), as well as children and adolescents (see review by McCutchen, 2011), to date, only a handful of studies have examined WM effects in L2 writing. Kormos and Sáfár (2008) investigated how Hungarian secondary school learners performed on writing tasks from a general language proficiency exam (Cambridge First Certificate Examination) and on tasks targeting PSTM and the simultaneous storage and processing function of WM. They found that, for those students who were at a pre-intermediate level of English, PSTM scores (i.e., performance on a non-word repetition test) correlated moderately and positively with their writing scores. In contrast, performance on the backward digit span task, which was used to assess the storage and processing function of WM, was not significantly associated with students' writing scores. Adams and Guillot (2008) studied 12 to 15-year old bilingual students' spelling and writing in French and English in relation to their PSTM, verbal and visual memory span. While no significant links between verbal or visual working memory and text composition emerged in either language, a significant relationship was found between PSTM scores and spelling in English.

More recently, Révész et al. (2017) investigated the relationship between adult L2 learners' WM functions and their writing processes and product when completing an opinion-writing task from an international language proficiency exam. Key-stroke logging, eye-tracking and stimulated recalls were used to capture participants' writing processes while composing the essays. This revealed that individuals with better task-switching ability paused for shorter periods between sentences, while those who had better ability to update information paused less frequently between paragraphs. Participants with smaller visual short-term memory capacity gazed at the instructions more often. Regarding text quality, a surprising finding was that those with higher PSTM scores used words from the first 1000 most frequent word-band more often, indicating a potentially negative role of the storage capacity of PSTM in lexical selection.

Zalbidea (2017) explored how WM mediates task complexity effects in oral and written L2 tasks in Spanish. In the written modality, only one significant correlation emerged with WM functioning: the number of errors against gender and plural markings on complex written tasks showed a strong negative correlation with WM scores. Zalbidea argued that more efficient WM functioning allows learners – even during complex tasks – to devote their limited attention to accuracy.

Zalbidea's (2017) study highlights that it is important to consider that the effects of WM on performance might depend on the complexity of the task learners perform. Robinson (2001) defines task complexity as "the result of attentional, memory, and other information processing demands imposed by the structure of the task on the language learner" (p. 29). Task complexity has been shown to affect spoken and written L2 performance (see meta-analysis by Jackson & Suethanapornkul, 2013). Robinson (2007) also argues that as task complexity increases, learners with more efficient WM functioning will be more successful in handling the growing cognitive and linguistic task demands. To date, very few studies have investigated interactions between task complexity and working memory in L2 performance. Kim, Payant and Pearson (2015) showed that L2 learners with a high level of WM functioning noticed more recasts during complex than simple oral tasks. Students with less efficient WM functioning were found to notice recasts less frequently regardless of the complexity of tasks. Zalbidea (2017), as mentioned, also revealed that only written performance on complex tasks was associated with WM test scores.

2.3. WM and L2 writing of young learners

The above review indicates that the role of WM functions in the L2 writing of older learners is far from conclusive. Even less is known about the impact of individual differences in WM on young L2 learners' writing processes and achievements. First, young learners' writing skills are still developing in their L1 throughout adolescence (Kellogg, 2008) because of the ongoing process of cognitive development. Through schooling, children also gain more expertise in writing as a technical skill (i.e., spelling, hand writing, keyboarding), in text composition and as a reader-oriented cognitive activity (Berninger, 1999; Isbell, 2017; MacArthur & Graham, 2016; McCutchen, 2011; Olive, 2012). Both cognitive development and experience are thought to have an impact on different aspects of text quality, such as syntactic complexity, lexical sophistication and discourse quality (Berninger, 1999; Kellogg, 2008; McCutchen, 2011). In addition, individual variation in WM functioning and differences among children in their cognitive developmental trajectories can put "a fundamental brake on the writing skill of developing writers throughout childhood, adolescence, and young adulthood" (Kellogg, 2008, p. 8; see also Berninger, 1999; Olive, 2012). Therefore, we can hypothesize that these maturational processes also affect L2 writing development. However, one of the few recent studies in this area did not find any significant links between the non-verbal intelligence and writing skills of young learners – either in L2 English or in L1 German in a bilingual schooling context (Steinlen, 2018).

Additionally, empirical work on the complex interactions of WM and task characteristics in the performance of young L2 learners is limited. Recent developments in the context of L2 assessment, however, have led to an increased usage of communicative goal-oriented task types such as integrated tasks that require a combination of multiple language skills (Cumming et al., 2005; Cushing-

Weigle, 2002; Wolf & Butler, 2017). These tasks require students to reproduce information from written or aural input in writing (Read-Write or Listen-Write tasks) or speaking (Read-Speak or Listen-Speak tasks), and thus might be demanding for individuals with less efficient WM functioning, particularly for young learners. As a result, students with low WM capacity might be disadvantaged in these integrated skills tasks when they are used in assessment – be it in high-stakes tests or classroom-based contexts – and the test might be unfairly biased towards those with better WM abilities. In other words, individual differences in WM functioning might create construct-irrelevant variance in integrated task performances, and the diagnostic and evaluative information gained about writing skills might be inaccurate. As young learners' cognitive abilities are still undergoing development, it is also possible that the role of WM functioning in L2 writing performance changes across grade levels. Therefore, it is imperative to investigate the role of WM in integrated tasks performed by younger learners for whom test results might determine their educational and professional future. Research on the effects of cognitive capacity limitations on various types of writing tasks might also yield useful information for teachers by helping them identify learners who might need additional support with certain types of tasks.

3. The present study: research questions

To the best of our knowledge, no earlier work has investigated young learners' performance on integrated tasks in relation to individual differences in WM functions. The present study aims to address this gap and uses performance data from several task types of a standardized computer-based test specifically designed for young L2 learners. Our study is also novel because it applies several tools to measure the different functions of WM (storage, processing, executive functioning) in order to provide insights into how cognitive capacity limitations might be related to performance differences. Furthermore, to examine whether writing experience and cognitive maturation might influence the relationship between cognitive functioning and writing test scores, we explored L2 writing in two different grade levels. More specifically, our study addressed the following research question:

RQ: What is the role of WM functioning, grade level and task type in the L2 writing performances of young English language learners?

Our hypotheses were that students' performances would differ across grade levels due to cognitive maturity and longer period of instruction. We also expected that task demands would exert a significant effect on students' writing scores and that children with more efficient WM functioning would achieve higher marks than those with less efficient WM functioning. It was further hypothesized that the role of WM functioning would be stronger in the lower grade and in tasks with higher formulation demand and which require the integration of listening and writing skills (cf. Kellogg, 2008).

4. Research methodology

The research was undertaken with young learners in Grades 6 and 7 (11–14 years old), in Hungary. They performed the Writing subsection of the TOEFL® Junior™ Comprehensive test-battery, which included four task types (Editing, Email, Opinion, Listen-Write). These tasks were designed by the test developers (Educational Testing Service – ETS) to reflect the most important writing domains that children might encounter in bilingual education contexts and to tap different writing processes (e.g., monitoring and composing). The four writing task types also vary in formulation demands. The Listen-Write task requires the rendering of given information, whereas in the Opinion task students need to express their own views. The Email task contains required elements of information as well as some optionality for individual ideas. The Editing task has no formulation demands as in this task students must detect and correct errors.

4.1. Participants

The participants were 94 L2 English learners in two primary schools in Budapest, Hungary. In both schools, which follow a bilingual education programme, children start learning English from Grade 1, have five English language classes per week all through primary school (up to Grade 8) and study a variety of subjects in English. The children receive content-based instruction through the medium of English in arts, music, science and physical education in lower primary school (Grades 2 to 4), and in history and science in upper primary school (Grades 5 to 8) – thus constituting a content-based language instruction (CLIL) context. Both schools are state-owned, and students receive education free of charge. Students learn approximately one third of the subjects (e.g., science, art, music and physical education) in English. The writing instruction the children received in the two schools was similar and very similar amounts of class time were devoted to writing activities.

Fifty-six children attended School A and 38 School B. Forty-five percent were boys and fifty-five percent girls. Their ages ranged between 11 and 14 years (M_{age} = 12.22, SD = .78). Fifty-four percent were enrolled in Grade 6 and 46 percent in Grade 7. All children had been learning English since Grade 1. Based on the students' overall performance on the full TOEFL® Junior™ Comprehensive test-battery, their English-L2 proficiency varied between the A2 to B2 level on the CEFR: 31 percent were at A2, 24 percent at B1, and 45 percent at B2 level.[1]

[1] An ANOVA test showed that the level of overall proficiency as measured by the TOEFL® Junior™ Comprehensive test did not differ significantly across schools ($F = 3.65$ $p = .06$) and grades ($F = 1.68$ $p = .19$) No interaction between school and grade was found either ($F = .38$ $p = .84$).

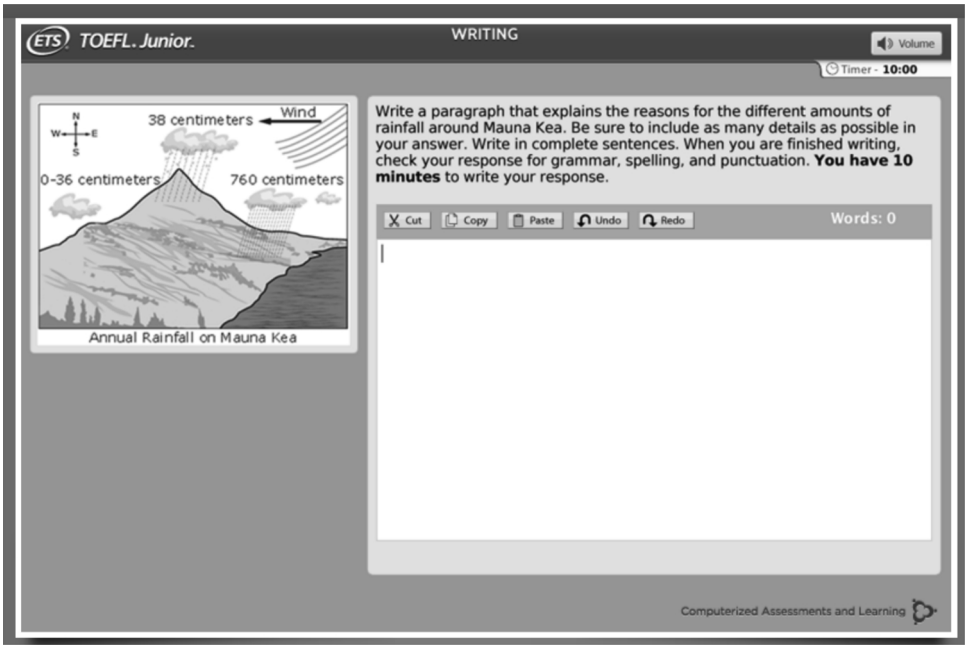

Fig. 1. Example integrated Listen-Write task. (Copyright© 2012 Education Testing Service – Reproduced with permission).

4.2. Instruments

4.2.1. Background questionnaire

To establish a demographic profile of the participant group, we designed a short bio-data questionnaire, asking participants about their gender, age, grade level, language(s) spoken at home, residence abroad, length of learning English and use of English outside the school context. The questionnaire was developed in English, translated into Hungarian and administered using the online survey tool Qualtrics.

4.2.2. Writing tasks

The pupils completed the computer-based TOEFL® Junior™ Comprehensive test battery, which involves reading, writing, listening, and speaking tasks in English. In this study, we examined performance on the writing section only. For reasons of test security, the exact content of the tasks cannot be revealed, but a general description of the four task types is as follows. The first type is an Editing task which requires test-takers to correct four errors in a paragraph of a non-academic and an academic text, respectively (Ed1/ Ed2). In the second one, an Email task, test-takers write a reply to an email. In the third one, an Opinion task, they compose a paragraph of 100–150 words in which they express their opinion on a topic. In the last one, an integrated Listen-Write task, test-takers first listen to a teacher talking about an academic topic while seeing a picture with animations, which include key information. The teacher's presentation lasts for about 90 s, and learners can take notes while listening. Test-takers are then asked to write a summary paragraph and check their responses for grammar and spelling. While typing the paragraph, the task instructions and illustration on the computer screen remain visible (see Fig. 1).[2]

4.2.3. Tasks measuring WM functioning

The participants also performed a series of WM tasks that take into account young learners' cognitive functions still under development (Gathercole et al., 2004). The specific WM tests were chosen based on three criteria: (1) language independence, (2) suitability for young learners, (3) feasibility regarding time restrictions of the research. The tasks aimed to measure the storage, processing and task-switching functions of WM. More specifically, to measure the storage and processing functions of WM, we used visual forward and backward digit span tasks. Digit span tests were originally developed for assessing intellectual abilities (IQ) in young children and are routinely used with children as young as four years old, as well as up to adulthood (see e.g., Gathercole et al., 2004; Jarvis & Gathercole, 2003). In our study, we chose visual over auditory span tasks or those using letters, (non-)words or sentences, because the former are seen to be less dependent on language than other versions. In addition, visual span tasks allowed us

[2] The time provided for the writing tasks was the same as the regular standardised limits for the TOEFL® Junior™ Comprehensive test and also showed to be sufficient for the participating students. The mean number of words written was within the required word-limit (Email task Mean = 78 words; Opinion task Mean = 110 words; Listen-Write task Mean = 111 words). However, it is important to note that text length is not considered in scoring students' performances.

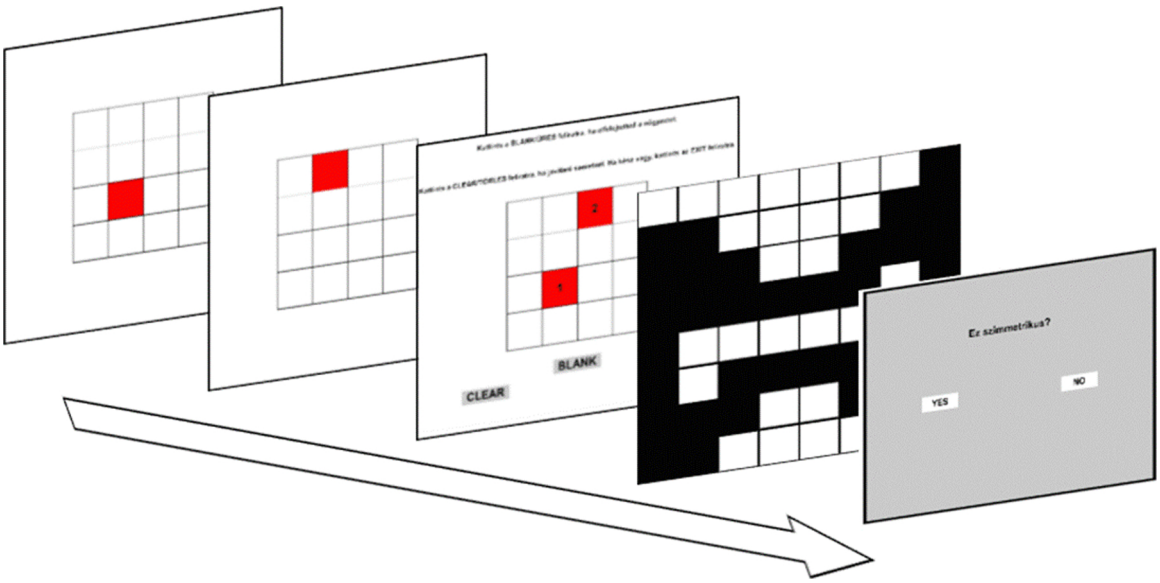

Fig. 2. Symmetry Span Task.

to test students in groups. Furthermore, in their review of WM tests for children and adolescents, Jarvis and Gathercole (2003) report that digit span tasks are suitable for the age group of 11–14-year-olds. We used digit span versions based on Woods et al. (2011; Experiment 1).

As a measure of the task-switching function of WM, we opted for the Symmetry Span task (SymSpan; Kane et al., 2004). The SymSpan task asks participants to remember the location of a sequence of blocks (e.g., in a 4 × 4 grid) while being interrupted by a decision task on the symmetry of a black-and-white block pattern (Conway et al., 2005; see Fig. 2).

The updating function of the WM was not measured in a separate test because recent research has shown large overlap between the updating and the processing functions of WM (Indrarathne & Kormos, 2018). Although we assessed students' inhibitory control with a Stop-Signal task (Logan, 1994), we do not report results relating to this task, since students' performance on this was highly positively skewed and the kurtosis value was also very high, indicating the presence of a large number of outliers. The Inquisit Web tool (www.millisecond.com) was used to set up and administer all WM tests.

4.3. Procedures

Ethical approval for the research was granted by the relevant ethics review committee at the researchers' institution, and consent was obtained from parents as well as the children. The WM and CE tasks and the TOEFL® Junior™ Comprehensive test were piloted with 14 students. Only minor rewordings of the Hungarian instructions were needed for the WM tasks. The piloting of the language test showed that the test was suitable for the children in terms of language proficiency level, structure and timing. The participants' perceptions of the test were also found to be positive. Furthermore, the children demonstrated an appropriate level of computer literacy and typing skills, and no technical issues arose during test performance. In both the pilot and the main study, the children's classroom teachers familiarized the learners with the test, using publicly available sample materials and the TOEFL® Junior™ Comprehensive test handbook.

Data collection took place in two consecutive sessions in Spring 2017. Participants first completed the WM tasks and then the TOEFL® Junior™ Comprehensive test. Finally, they completed a short online bio-data questionnaire. All instruments were group administered, with one of the authors and two research assistants overseeing the procedures. The participants spent 25–35 minutes in total on the WM tasks: the visual forward and backward digit span tasks took about 5 min each and the SymSpan and Stop-Signal tasks about 10 min each. All TOEFL® Junior™ Comprehensive tasks were computer-administered. The Editing tasks had a time limit of 2.5 min each. Learners had 7 min for composing in the Email task, 10 min for the opinion task, and 10 min to write a summary in the

Table 1
Overview of writing tasks.

	Editing tasks	Email task	Opinion task	Listen-Write task
Task	Correcting four errors in an academic and non-academic text	Read an email and write a response email	Write a 100-150 word paragraph to express opinion	Summarize a lecture on an academic topic
Length	5 minutes	7 minutes	10 minutes	10 minutes

Listen-Write task (see Table 1 for an overview). Throughout the experiment, the learners were given several breaks (in line with the regulations of the TOEFL® Junior™ Comprehensive test).

4.4. Scoring and analysis

Experienced raters from the TOEFL® Junior™ Writing rater pool scored the pupils' performances on the writing tasks based on the TOEFL® Junior™ Comprehensive performance descriptors (https://www.ets.org/s/toefl_junior/pdf/toefl_junior_comprehensive_writing_scoring_guides.pdf).[3] Accordingly, participants could achieve a maximum score of 4 on each part of the writing section, where the top level indicates that the writer produced an accurate and coherent text by using simple and complex sentences to provide key information. A top score also demonstrates that the test-taker understood and accurately conveyed key ideas and sufficient supporting detail. A score of zero represents either no response or a response that is off-task. The maximum total score participants could achieve was 4 on each task, i.e., 16 in total (scores on the two Editing tasks were averaged).

Following Woods et al. (2011), the Forward and Backward Digit Span scores provide an estimate of the score each participant was expected to get correct 50 percent of the time based on overall performance during all 14 trials. The SymSpan score gives the sum of all accurately recalled items in a correct order (Conway et al., 2005).

To investigate the relationships between the predictor variables and the TOEFL Writing scores, we used CLMMs (Christensen, 2015; Agresti, 2010), also known as multilevel ordinal regression models, to analyse the 470 observations – 94 students completing five writing tasks each – using the clmm function in the Ordinal package (Christensen, 2015) in R (R Core Team, 2018). CLMMs were appropriate for two reasons. First, we had a crossed random effect: each student completed five writing tasks. Second, the outcome variable (writing performance) can be considered an ordinal scale whereby there is ordering of the levels (from 0 to 4) and an upper (4) and lower (0) limit for each writing task. CLMMs allow us to account for the potential ceiling and floor effects imposed by these limits. The predictor variables included: Grade (Grade 6 vs Grade 7), Writing Task (Task 1 Edit 1, Task 1 Edit 2, Task 2 Email, Task 3 Opinion, Task 4 Listen-Write), and a composite score of WM (see next section for details).

5. Results

5.1. Descriptive statistics and correlational analyses

In Table 2 the descriptive statistics of the different writing task performances are given. Accordingly, participants performed particularly well on the Email and Listen-Write tasks with mean scores above 3 out of 4. Students also scored high on the Opinion task with a mean score of approximately 3. The two Editing tasks received lower scores.

Table 3 provides an overview of the descriptive statistics for the WM data. Despite their young ages, our participants achieved high Forward and Backward Digit Span scores of 6 and 5.5, respectively. In comparison, Jarvis and Gathercole (2003) obtained scores of 5.2 and 4.6 for 14 year olds, while the adolescents in Kormos and Sáfár (2008) achieved 5.3 on the backward digit span. Participants reached a task switching score (SymSpan task) of almost 19. Standard deviations and minimum and maximum scores indicate that the young learners in our study displayed a wide range of cognitive abilities.

Correlational analyses (Table 4) between the three different WM tests showed that the forward and backward versions of the Digit Span test assessed a partly overlapping construct, as indicated by the strong correlation between the two measures. The Symmetry Span test correlated moderately with the two-digit spans.

To minimise the risk of Type I and Type II errors due to potential multicollinearity (Tu et al., 2005), we explored whether it would be appropriate to create a composite score of the three WM test scores. A principal component analysis with the three test scores obtained a Kaiser-Meyer-Olkin measure of sampling adequacy of .64. This lies above the recommended minimum value of .50 (Pett et al., 2003). A Bartlett's Test of Sphericity (Approximate $\chi^2(3) = 73.70$, $p < .001$) also indicated that the correlation matrix could be combined into one factor. The total variance table revealed that, together, the span scores had an Eigenvalue of 1.95, which could explain 65.10% of the variance with respective initial Eigenvalues of 65% (ForwardDigitSpan), 22% (BackwardDigitSpan) and 13% (SymSpan). The factor loadings showed that the three components contributed to the score to a similar extent: ForwardDigitSpan = .83; BackwardDigitSpan = .87; SymSpan = .72. Based on these commonalities among the three span tasks, and the results of the factor analysis, we deemed it appropriate to create a composite score using regression factor scores (Tabachnick & Fidell, 2001).

5.2. The role of WM in writing scores

To answer our research question, we started by fitting a series of models, beginning with a minimal model containing just the random effects of students on intercepts, and progressively increasing the model complexity by adding fixed effects and interaction terms. The minimal model (Model 1) was compared to a model including terms corresponding to the fixed effects of: Grade Level, Task and WM (Model 2). The Likelihood Ratio Tests (LRT) revealed that the additional complexity of the model was justified. Model 2 provided a better fit for the data than Model 1, $\chi^2(6) = 190.32$, $p < .01$. Next, we compared Model 2 to a model with added interactions of: Grade Level by Task; Grade Level by WM; Grade Level by Task by WM (Model 3). We found that the inclusion of

[3] ETS employs strict quality control of rating procedures and rater monitoring. More information can be found at: https://www.ets.org/scoring_opportunities/.

Table 2
Descriptive statistics of the different writing task performances (N = 94).

	Task 1 Ed1	Task 1 Ed2	Task 2 Email	Task 3 Opinion	Task 4 Listen-Write
Mean	1.94	2.24	3.24	2.81	3.13
SD	1.08	1.14	.76	.82	.81
Minimum	0	0	0	1	0
Maximum	4	4	4	4	4
Skewness (*SE*)	.08 (.25)	−.10 (.25)	−1.05 (.25)	.01 (.25)	−.87 (.25)
Kurtosis (*SE*)	−.71 (.49)	−.92 (.49)	2.20 (.49)	−.85 (.49)	1.27 (.49)

Note. *SE* = Standard Error

Table 3
Descriptive Statistics on WM tests (N = 94).

	ForwardDigitSpan[a]	BackwardDigitSpan[a]	SymSpan[b]
Mean	6.08 (.09)	5.58 (.09)	18.83 (.82)
SD	.91	.93	8.44
Minimum	4.39	3.79	2
Maximum	9.25	8.25	42
Skewness (*SE*)	.41 (.25)	.37 (.25)	.58 (.25)
Kurtosis (*SE*)	.66 (.49)	.47 (.49)	−.02 (.49)

Note. [a] Score that a participant is expected to get correct 50 percent of the time based on overall performance during all 14 trials; [b] Sum of all correctly recalled squares of correctly recalled sets; *SE* = Standard Error.

Table 4
Pearson correlations between WM tests (N = 94).

	BackwardDigitSpan	SymSpan
ForwardDigitSpan	.60**	.43***
BackwardDigitSpan		.52***

Note. ** $p < .01$; *** $p < .001$.

interactions further improved the model fit, $\chi^2(13) = 22.46$, $p < .05$. Thus, based on both the theoretical interest associated with interactions and improvement in the model fit to the data, we decided to keep the interactions.

In the next step of the analysis we found that the Maximum Likelihood Model was too complex for the underlying data. The random effects structure exceeded the number of observations in the data set; therefore, the model did not converge as it was over-parameterised. Following the recommendation of Bates, Kliegl, Vasishth, and Baayen (2015), to keep the model parsimonious, we established the utility of random slopes using the LRT. We found that the addition of random slopes did not improve the model fit. Consequently, our final model contained the random intercept of students only.

A summary of the final model is presented in Table 5 where we supplement the log-odds estimates with Odds Ratio (OR) estimates. It is important to mention that the summary table of the final model should not be interpreted directly, as all the coefficients are estimated against the reference level categories. For example, the significant coefficient for Grade 7 indicates that Grade 7 students were 3.86 times more likely to obtain a higher writing score than Grade 6 students, but only for the reference level of the writing task (Task 4 Listen-Write) and keeping WM at the average. For a more intuitive interpretation of estimates, subsequent multiple comparisons tables should be considered (Tables 6–9), where we adjusted the *p*-value for multiple comparisons using normal approximation. Expected Score Change estimates in Tables 6–8 demonstrate the impact of the model's estimates on expected writing scores for the subgroup being compared. This allows for a more meaningful interpretation of the estimates and we recommend considering both the OR and Expected Score Change estimates alongside the significance values. For example, in Table 7 we can see that, keeping WM at the average, Grade 7 students were 3.86 times more likely to score higher on Task 4 Listen-Write than Grade 6 students, and although this difference is significant on average it is not expected that Grade 7 students will have a higher score on that task than Grade 6 students. The expected writing scores were calculated by considering both the model estimates for between-predictor comparisons and the four intercept parameters which show the log-odds thresholds that specify the expected writing score.

To look into the role of task type and grade level in L2 writing performances, we investigated how tasks varied across grade levels and WM (Table 6). Keeping WM constant at a *z*-score of zero, both grade levels were on average more likely to have a higher writing score on Task 2 Email versus every other task except Task 4 (Listen-Write). The biggest difference for both Grade levels was between Task 2 Email and Task 1 Ed1 where Grade 7 students were 40.85 times (Grade 6 were 37.71 times) more likely to score higher on Task 2 Email than Task 1 Ed1, and the expected writing score change for Grade 7 was from 2 (Task 1 Ed1) to 4 (Task 2 Email). However, in some cases the significant difference between Task 2 Email and other tasks did not result in any meaningful expected

Table 5
Summary of the Writing Model.

	Estimate	OR	Standard Error	z-value	p
Grade 7	1.35	3.86	.49	2.73	**
Task 1 Ed1	−2.97	.05	.42	−7.15	***
Task 1 Ed2	−1.90	.15	.40	−4.73	***
Task 2 Email	.66	1.93	.39	1.69	.09
Task 3 Opinion	−.54	.58	.38	−1.41	.16
zWM	.31	1.36	.36	.88	.38
Grade 7 x Task 1 Ed1	−.33	.72	.58	−.57	.57
Grade 7 x Task 1 Ed2	−.73	.48	.58	−1.25	.21
Grade 7 x Task 2 Email	−.25	.78	.60	−.41	.68
Grade 7 x Task 3 Opinion	−.85	.43	.57	−1.49	.14
Grade 7 x zWM	.09	1.09	.49	.18	.86
Task 1 Ed1 x zWM	−.11	.90	.42	−.27	.79
Task 1 Ed2 x zWM	1.10	3.00	.44	2.52	*
Task 2 Email x zWM	.15	1.16	.43	.35	.72
Task 3 Opinion x zWM	−.01	.99	.41	−.02	.98
Grade 7 x Task 1 Ed1 x zWM	.23	1.26	.57	.41	.68
Grade 7 x Task 1 Ed2 x zWM	−.10	.90	.59	−.17	.86
Grade 7 x Task 2 Email x zWM	−.31	.73	.59	−.53	.59
Grade 7 x Task 3 Opinion x zWM	.13	1.14	.57	.23	.82
Random Intercept	Variance		Standard Deviation		
Students	1.61		1.27		
Score Thresholds	Estimate		Standard Error	z-value	
0|1	−5.62		.49	−11.56	
1|2	−3.32		.39	−8.59	
2|3	−1.05		.34	−3.12	
3|4	1.64		.35	4.73	

Note 1. * = $p < .05$; ** = $p < .01$; *** = $p < .001$. *Note 2.* Grade 6 is the reference level for Grade; Task 4 Listen-Write is the reference level for Task. *Note 3.* zWM is centred and standardised WM. *Note 4.* OR refers to Odds Ratio.

score change. For example, students in Grade 6 were 3.33 (1/.30) times less likely to score higher in Task 3 Opinion than in Task 2 Email, but for both tasks the average expected writing score is 3. This contradiction occurs because the position of the baseline group (Task 3 Opinion) at the lower end of the threshold 3 intervals is such that the effect of the significant but small OR change means that the resulting score remains in threshold 3 for the comparison group (Task 2 Email). On the other hand, students in Grade 7 were 1.52 times more likely to score higher in Task 2 Email versus Task 4 Listen-Write, and although the difference in log-odds between the two tasks is not significant, it resulted in an expected writing score change from 3 to 4, whereby Grade 7 students were expected to score 4 on Task 2 Email but only 3 on Task 4 Listen-Write. In this case, this happens because the baseline scenario is close to the 3|4 threshold and does not require an increase in OR of significant magnitude to cross that boundary; we describe this as a meaningful effect.

With one z-score increase in WM, students in both Grades were more likely to score highest on Task 2 (Email) versus any other task. However, in most cases for Grade 6 students, the differences between Email writing and other tasks resulted in no change in the expected writing score. Task 1 Ed1 was the only task among the Grade 6 students which had an expected writing score of 2. Consequently, those with higher WM in Grade 6 were significantly more likely to score lower on Task 1 Ed1 than on any other task. For example, students in Grade 6 were 49.40 times more likely to score higher on Task 2 Email than Task 1 Ed1. Students with higher WM in Grade 7 also scored significantly lower in Task 1 Ed1 than on any other task, with the most meaningful change being between Task 1 Ed1 (expected writing score 2) versus Task 4 Listen-Write and Task 2 Email (both had expected writing scores of 4). Overall, the number of significant differences between the tasks reduced with a single z-score increase in WM, but this did not reduce the number of meaningful differences between the tasks among Grade 7 students.

Table 7 indicates that from Grade 6 to Grade 7 the writing skills have improved in terms of log-odds. However, in most tasks, including Task 4 Listen-Write, on average it is not enough to see a detectable change in the average writing score. Interestingly, Grade 7 students were 3 times more likely to score higher on Task 2 Email than Grade 6 students, and although this difference is not significant, it is meaningful since it results in an expected score change from 3 for Grade 6 to 4 for Grade 7.

Our comparisons also revealed that students with above average WM had higher writing score log-odds on Task 1 Ed2 than those with lower WM (Table 8). Grade 6 students with a WM score of one standard deviation higher than average students were 4.10 more likely to have a higher writing score on Task 1 Ed2 than the students with average WM, whereas their Grade 7 counterparts were 4.06 times more likely to have a higher writing score on this task. Additionally, the influence of WM on Task 1 Ed2 was meaningful since students in both grades with above average WM performance were expected to score 3, whereas those with average WM performance were expected to score 2. Interestingly, WM had a non-significant, but meaningful, effect on Task 4 Listen-Write scores among Grade 7 students, whereby those with higher WM were 1.49 times more likely to have a higher writing score (expected score of 4) than those with lower WM (expected score of 3). Table 9 demonstrates that an increase in WM resulted in a marginally better improvement in log-odds for one Grade level versus the other, but this was not significant. We do not show the expected score change in Table 9 since the ORs here describe changes in strength of WM influence between two participant groups rather than its influence on a specific group as evaluated earlier.

Table 6
Multiple comparisons: Tasks by Grade.

Comparison	Estimate	OR	Expected Score Change	Standard Error	z-value	p
Grade 6†						
Task 1 Ed1 - Task 4 Listen-Write	−2.97	.05	3 → 2	.41	−7.28	***
Task 1 Ed2 - Task 4 Listen-Write	−1.90	.15	3 → 2	.40	−4.78	***
Task 2 Email - Task 4 Listen-Write	.66	1.93	3 → 3	.39	1.69	.44
Task 3 Opinion - Task 4 Listen-Write	−.54	.58	3 → 3	.38	−1.41	.62
Task 1 Ed2 - Task 1 Ed1	1.07	2.92	2 → 2	.39	2.75	*
Task 2 Email - Task 1 Ed1	3.63	37.71	2 → 3	.42	8.61	***
Task 3 Opinion - Task 1 Ed1	2.43	11.36	2 → 3	.40	6.10	***
Task 2 Email - Task 1 Ed2	2.56	12.94	2 → 3	.41	6.26	***
Task 3 Opinion - Task 1 Ed2	1.36	3.90	2 → 3	.39	3.48	**
Task 3 Opinion - Task 2 Email	−1.20	.30	3 → 3	.39	−3.07	*
Grade 7†						
Task 1 Ed1 - Task 4 Listen-Write	−3.30	.04	3 → 2	.45	−7.32	***
Task 1 Ed2 - Task 4 Listen-Write	−2.63	.07	3 → 2	.45	−5.89	***
Task 2 Email - Task 4 Listen-Write	.42	1.52	3 → 4	.45	.93	.89
Task 3 Opinion - Task 4 Listen-Write	−1.39	.25	3 → 3	.43	−3.27	**
Task 1 Ed2 - Task 1 Ed1	.67	1.95	2 → 2	.43	1.55	.53
Task 2 Email - Task 1 Ed1	3.71	40.85	2 → 4	.47	7.92	***
Task 3 Opinion - Task 1 Ed1	1.91	6.75	2 → 3	.42	4.51	***
Task 2 Email - Task 1 Ed2	3.05	21.12	2 → 4	.46	6.57	***
Task 3 Opinion - Task 1 Ed2	1.24	3.46	2 → 3	.42	2.94	*
Task 3 Opinion - Task 2 Email	−1.81	.16	4 → 3	.44	−4.08	***
Grade 6‡						
Task 1 Ed1 - Task 4 Listen-Write	−3.08	.05	3 → 2	.61	−5.08	***
Task 1 Ed2 - Task 4 Listen-Write	−.80	.45	3 → 3	.62	−1.29	.70
Task 2 Email - Task 4 Listen-Write	.81	2.25	3 → 3	.62	1.32	.68
Task 3 Opinion - Task 4 Listen-Write	−.55	.58	3 → 3	.60	−.92	.89
Task 1 Ed2 - Task 1 Ed1	2.28	9.78	2 → 3	.62	3.71	**
Task 2 Email - Task 1 Ed1	3.90	49.40	2 → 3	.63	6.23	***
Task 3 Opinion - Task 1 Ed1	2.53	12.55	2 → 3	.59	4.28	***
Task 2 Email - Task 1 Ed2	1.61	5.00	3 → 3	.63	2.54	.08
Task 3 Opinion - Task 1 Ed2	.25	1.28	3 → 3	.61	.41	.99
Task 3 Opinion - Task 2 Email	−1.36	.26	3 → 3	.61	−2.24	.16
Grade 7‡						
Task 1 Ed1 - Task 4 Listen-Write	−3.18	.04	4 → 2	.57	−5.55	***
Task 1 Ed2 - Task 4 Listen-Write	−1.63	.20	4 → 3	.57	−2.89	*
Task 2 Email - Task 4 Listen-Write	.25	1.28	4 → 4	.58	.44	.99
Task 3 Opinion - Task 4 Listen-Write	−1.27	.28	4 → 3	.55	−2.31	.14
Task 1 Ed2 - Task 1 Ed1	1.54	4.66	2 → 3	.56	2.75	*
Task 2 Email - Task 1 Ed1	3.43	30.88	2 → 4	.59	5.81	***
Task 3 Opinion - Task 1 Ed1	1.91	6.75	2 → 3	.55	3.46	**
Task 2 Email - Task 1 Ed2	1.89	6.62	3 → 4	.58	3.24	*
Task 3 Opinion - Task 1 Ed2	.36	1.43	3 → 3	.55	.67	.96
Task 3 Opinion - Task 2 Email	−1.52	.22	4 → 3	.57	−2.69	.06

Note 1. * = $p < .05$; ** = $p < .01$; *** = $p < .001$. *Note 2.*† = keeping WM constant at a z-score of 0; ‡ = after an increase in WM by a z-score of 1.

Table 7
Multiple comparisons: Grade by Task keeping WM constant (z-score = 0).

Comparison	Estimate	OR	Expected Score Change	Standard Error	z-value	p
Gr_7 Task 1 Ed1 - Gr_6 Task 1 Ed1	1.02	2.77	2 → 2	.48	2.11	.15
Gr_7 Task 1 Ed2 - Gr_6 Task 1 Ed2	.62	1.86	2 → 2	.49	1.26	.65
Gr_7 Task 2 Email - Gr_6 Task 2 Email	1.10	3.00	3 → 4	.51	2.17	.13
Gr_7 Task 3 Opinion - Gr_6 Task 3 Opinion	.50	1.65	3 → 3	.47	1.06	.79
Gr_7 Task 4 Listen-Write - Gr_6 Task 4 Listen-Write	1.35	3.86	3 → 3	.49	2.74	*

Note 1. * = $p < .05$.

Table 8
Multiple comparisons: increase in WM (z-score = 1) by Task and Grade.

Comparison	Estimate	OR	Expected Score Change	Standard Error	*z*-value	*p*
Grade 6						
Task 1 Ed1	.20	1.22	2 → 2	.35	.57	.98
Task 1 Ed2	1.41	4.10	2 → 3	.37	3.82	***
Task 2 Email	.46	1.58	3 → 3	.36	1.30	.62
Task 3 Opinion	.30	1.35	3 → 3	.34	.89	.88
Task 4 Listen-Write	.31	1.36	3 → 3	.36	.88	.89
Grade 7						
Task 1 Ed1	.52	1.68	2 → 2	.32	1.60	.41
Task 1 Ed2	1.40	4.06	2 → 3	.34	4.07	***
Task 2 Email	.24	1.27	4 → 4	.35	.68	.96
Task 3 Opinion	.52	1.68	3 → 3	.32	1.60	.41
Task 4 Listen-Write	.40	1.49	3 → 4	.33	1.20	.69

Note 1. *** = $p < .001$.

Table 9
Multiple comparisons: how log-odds change in Grade 7 against Grade 6 for a one unit increase in WM per task.

Comparison	Estimate	OR	Standard Error	*z*-value	*p*
Gr_7 Task 1 Ed1 - Gr_6 Task 1 Ed1	.32	1.38	.48	.67	.96
Gr_7 Task 1 Ed2 - Gr_6 Task 1 Ed2	−.02	.98	.50	−.03	1.00
Gr_7 Task 2 Email - Gr_6 Task 2 Email	−.23	.79	.50	−.46	.99
Gr_7 Task 3 Opinion - Gr_6 Task 3 Opinion	.21	1.23	.47	.45	.99
Gr_7 Task 4 Listen-Write - Gr_6 Task 4 Listen-Write	.09	1.09	.49	.18	1.00

6. Discussion

This study set out to investigate the role of WM functioning, grade level, and task type in L2 writing performances of young English language learners. We also aimed to establish whether WM functions play a differential role in L2 writing performance depending on grade level and task type.

Our results showed that the participants performed well on the writing tasks of the computer-administered TOEFL® Junior™ Comprehensive test. After nearly six and seven years of intensive English learning (5 h per week language classes plus approximately 5–7 h of content-based instruction per week) respectively, they reached A2 to B1 level, some even performing at B2 (cf. Tannenbaum & Baron, 2015, p. 16). This shows that the test tasks were within the competence of the targeted sample of young learners in the investigated CLIL context. It was also encouraging to see that participants did particularly well on the Email and Listen-Write tasks where, on average, they scored above 3 out of 4 points. The fact that the integrated task type (Task 4 Listen-Write) elicited similarly high scores as the Email and Opinion tasks, but that the two Editing tasks (Task 1 Ed 1/2) significantly differed from the other task types, gives further support to earlier calls to create and apply instructional and assessment tasks that reflect target language use domains in academic contexts (Cumming et al., 2005; Cushing-Weigle, 2002; So et al., 2015) (see below for more discussion on the nature of the Editing task type). It also provides empirical evidence for the value of recent endeavours in L2 test design to assess test-takers by means of integrated task types that combine multiple language skills (So et al., 2015). This authentic integrated task type seems to present an appropriate-level challenge to L2 learners in a CLIL context and may prove to have high instructional value.

The descriptive statistics for the WM and CE test scores revealed that the young learners in our study, aged 11–14 years, achieved relatively high scores on the forward and backward digit span tasks (around 6 and 5.5, respectively) (see Jarvis & Gathercole, 2003, and Kormos & Sáfár, 2008, for comparisons). The mean task switching score (SymSpan task) in this group was almost 19, which is a lower capacity than that typically found in adult samples (cf. Foster et al.'s, 2015, mean score of 26.6 for healthy adults).

Given the range of scores in the WM and CE tests, it was unexpected to find that WM functioning had a limited effect on the L2 writing scores of the young learners, except for the academic version of the editing task (Task 1 Edit 2) and the integrated Listen-Write task in Grade 7. The lack of association between WM functioning and performance on most of the writing tasks in this study is surprising because WM functions have been assumed to influence the coordination, parallel processing of information and switching between sub-tasks of text composition (Olive, 2012; Révész et al., 2017). Our results, however, are partially in line with earlier research findings with older learners where significant relationships between writing scores and WM emerged only for the storage function of PSTM (Adams & Guillot, 2008; Kormos & Sáfár, 2008), but where limited or no relationships were found for the simultaneous storage and processing function of WM or executive control (Révész et al., 2017; Steinlen, 2018; Zalbidea, 2017). On the one hand, the findings of this study might show that varying levels of WM functions do not seem to cause construct-irrelevant variance in most of the writing tasks of the TOEFL® Junior™ Comprehensive test. Therefore, these types of tasks may serve as accurate tools for teachers and other stakeholders for gaining information about young learners' L2 writing skills in a CLIL context regardless of differences in WM functioning. On the other hand, the results might also indicate that the instruction the participants received in

their current CLIL context might have been beneficial in reducing variance in L2 writing skills that can potentially be caused by differential WM functioning among children.

One of the significant relationships between WM and writing performance was found in the Editing task where students had to find and correct errors in an academic text (Task 1 Edit 2) but not in the version that contained a non-academic text. This finding might corroborate the interpretation of Zalbidea (2017), who concluded that more efficient WM functions allowed her participants to devote more attention to accuracy, which was also an important aspect of the editing task in our study. In fact, in several other test systems, editing tasks are classified as language-in-use tasks (see e.g., the Austrian school-leaving exam for foreign languages (Froetscher, 2016). The Editing task, which involves metacognitive processing, assesses a specific aspect of the writing process, which is strongly inter-related with reading, namely the monitoring stage (Kellogg, 1996). The fact that WM effects were found for the academic editing task, but not for the non-academic version of the task, suggests that participants with more efficient WM functioning could have been more successful in co-ordinating their reading and monitoring processes when reading and monitoring accuracy in an academic text. Similar WM effects when reading complex texts have been detected in studies with monolingual children. These studies found that comprehension monitoring ability, which is often assessed with editing tasks and similar to those in our research, is related to WM functioning (e.g., Oakhill, Hart, & Samols, 2005). These results suggest that teachers might need to provide more assistance to young L2 learners with less efficient WM functioning to detect errors in their writing.

Another interesting finding of our study was the non-significant, but meaningful influence of WM functioning on performance in the Listen-Write task in Grade 7. The Listen-Write task can be considered a relatively complex task type, as it requires young learners to recall and summarize the content of aural input with support from visual input. Therefore, it is possible that the ability to efficiently coordinate attentional processes assists young learners in successfully executing these listening and writing processes. What is surprising, however, is that we found that WM functioning only played a meaningful role in Grade 7. We would have expected performance in this integrated task would be more prone to WM effects in the lower Grade 6 (cf. Kellogg, 2008). The reasons for the expected score change can probably be attributed to the fact that the expected score of students with average WM was closer to the highest score for Grade 7 than Grade 6. Thus, Grade 7 students did not need to improve as much as Grade 6 students to cross the threshold into a higher score. Nonetheless, in assessment contexts, especially if they are high-stakes, this score change might have an effect on the final grade, especially if only few tasks are used.

The influence of an increase in WM values on writing score log-odds was found to be similar for both grade levels. It should be noted that there was some variation and overlap in participants' ages within grade levels and the sample size of the grade level groups was relatively small, which might also account for this finding. However, a more likely explanation is that the one grade level difference in CLIL schooling in early adolescent years might not impact on how WM functioning influences L2 writing achievement.

Regarding differences across grade level, the analysis showed that, on average and irrespective of WM functions, participants in Grade 7 consistently outperformed Grade 6 learners, but this difference was only significant in Task 4 (Listen-Write). This difference, however, was not detectable as a change in the expected average writing scores. The fact that young learners with an added year of writing experience were found to perform somewhat better on this task might indicate that they have more-developed skills of coordinating the simultaneously ongoing operations that are necessary to perform successfully on this task type (Kellogg, 2008; Olive, 2012). The results also show a meaningful, albeit not significant difference between Grade 7 and Grade 6 students in the Email task. The one additional year of language instruction and cognitive maturity might have assisted Grade 7 students in this task type which was also relatively demanding in that it required comprehending the writing prompt (email) and responding with appropriate information.

This study also aimed to explore how task type might influence the writing performance of young learners and how WM functions might mediate task type effects. When WM scores were kept constant, students achieved significantly lower scores on the editing task than on the other types of tasks. In Grade 6, significant differences in the academic and non-academic versions of the editing tasks were also detected. When WM values were increased by one *z*-score, we found that both Grade 6 and Grade 7 learners scored significantly lower on the non-academic versions of the editing task than on other types. In this analysis, which models performance with high WM abilities, the students also received lower scores for editing a non-academic text than an academic one. As the test included only one non-academic and one academic editing task, it is difficult to attribute these findings to the academic nature of the text or to the specific type of the task. However, from the perspective of cognitive validity[4] (O'Sullivan & Weir, 2011), the use of the editing task in the assessment and teaching of young L2 learners, especially if the text contains academic content, needs to be carefully considered.

A further task-related difference was that when WM scores were kept at average, both Grade 6 and Grade 7 participants scored significantly lower on the task where they had to elaborate their opinion than on the email writing task. Nevertheless, the difference was only meaningful in the sense that it resulted in an expected grade change in Grade 7. Although the difference between these two tasks was not statistically significant for participants with above-average WM scores in Grade 7, we detected a meaningful change between the scores. The opinion-writing task differs from the email task in that it requires students to formulate their own ideas, whereas in the email task some of the content is specified in the prompt. This difference in formulation demands and students' potential lack of experience of writing in the argumentative genre in the Hungarian context might explain the lower performance of our participants on the opinion task. In fact, work by Olive (2012) and colleagues suggests that the maturation of cognitive effort in

[4] Cognitive validity concerns the extent to which the cognitive processes required to complete a task are appropriate, and likely to be used if the task was performed in the 'real world'. The socio-cognitive model also directly relates this to who the test taker is: "individual characteristics will directly impact on the way the individuals process the test task" (Weir, 2005, p. 51).

writing might become specifically visible on argumentative tasks. Indeed, the difference in scores seems to widen with age and exposure to L2, which suggests that young L2 writers would benefit from more practice in expressing their own opinion in writing and from more explicit instruction in argumentative genres across the curriculum.

7. Conclusions and implications

This study investigated the writing performances of young English-L2 learners and their relationship with individual differences in WM functioning. We also examined how L2 writing achievement might relate to task type and grade level and whether the effect of WM functioning on L2 writing performances varies across different types of tasks and years of study. Our research showed that young L2 learners performed particularly well on the Email writing and integrated Listen-Write tasks. When compared to Grade 6 students, Grade 7 participants scored statistically significantly higher on the Listen-Write task and meaningfully higher on the Email task. Both of these tasks are cognitively demanding as the integrated Listen-Write task requires writers to summarize information they had previously heard, and in the Email task students have to understand the original message in the email and respond to it. Differences in WM functioning played a limited role when young learners completed the written component of the TOEFL® Junior™ Comprehensive test-battery. The only task where statistically significant WM effects were detected was the academic editing task and a meaningful but non-significant WM influence was found for Grade 7 students in the Listen-Write task. The results also revealed that learners with high WM functions showed somewhat more consistent performance across tasks than did learners with low WM functions.

The findings of the present study have valuable implications for the practice of teaching and assessing L2 writing for young learners. First, the fact that individual differences in WM functions showed only limited interactions with scores on the writing section of the computer-administrated TOEFL® Junior™ Comprehensive test-battery provides support for the cognitive validity of this assessment tool. These findings indicate that, except for the academic editing task and potentially the Listen-Write task, young learners with less efficient WM functioning do not seem to be disadvantaged in the written tasks of the test. This finding is important for assessment and instructional task design, because it shows that students who have below average WM functions – and this group might potentially include learners with specific learning difficulties (cf. Kormos, 2017) – can perform on these tasks to the best of their knowledge, even under standard test administration and teaching conditions. In combination with our earlier findings that the same group of L2 learners also demonstrated positive attitudes and task motivation towards this test (Kormos, Brunfaut, & Michel, submitted), these results suggest that the writing tasks of the TOEFL® Junior™ Comprehensive test are appropriately tailored to the characteristics of young L2 learners (Bailey, 2017) and can yield useful diagnostic information on the writing development of young L2 learners in CLIL contexts. The results of our study might also contribute to positive washback effects on the teaching of writing for these students (Cheng, Watanabe, & Curtis, 2004). If standardized tests for young learners assess performance on tasks sampling from the target language use domain (e.g., informal and formal school interactions) and task types reflecting that domain (e.g., email writing or integrated tasks such as the Listen-Write task), it is hoped that material designers and teachers will also use these tasks more frequently in coursebooks and in the classroom.

This study is not without its limitations, however. First, it should be noted that we investigated a specific group of young learners that attended bilingual schools in Hungary. Data collection took place in the students' school environment and the assessment scores were only made available to the students and their parents, not to their teachers, and did not count towards their academic grades at school. Therefore, our sample and the conditions in which the test was administered might not be representative of the wide variety of young L2 learners and contexts in which students take the TOEFL® Junior™ Comprehensive test-battery or similar young learners tests.

A potential further limitation is the fact that we analysed our data using a composite score for the different WM tasks. However, this was essential for modelling, since we wanted to minimise the risks of Type I and Type II errors which are associated with multicollinearity. Although correlational analyses with individual WM tests, which we do not report in this paper, did not reveal a different pattern of relationships, in the future it would be interesting to explore interrelations of young L2 learners' writing scores with the individual components of WM to test the predictions of Kellogg's (1996) model of the writing process. Future work could also follow recent endeavours in researching adult L2 writing from a process-oriented perspective (for example Révész, Michel, and Lee (2019) who used key-stroke logging, eye-tracking, and stimulated recall). Another potential direction for future research could be the exploration of how cognitive maturation and experience (Berninger, 1999; McCutchen, 2011) might change the writing product and process of children and adolescents. Finally, replications of this project could be conducted using similar types of tasks with different content, as in our study – except for the editing task – each task type was only represented by one task.

Acknowledgements

This research was funded by Educational Testing Service (ETS), USA under a Committee of Examiners and TOEFL research grant. ETS does not discount or endorse the methodology, results, implications or opinions presented by the researcher(s).

We would like to thank all the young learners who took part in the study, as well as their parents, teachers and schools who supported the project with great enthusiasm. We owe our thanks to Dr. Simon Taylor at the Department of Mathematics and Statistics of Lancaster University for his help with the statistical analyses and to Dr. Gabriella Dóczi Vámos, Stella Varga and Orsolya Szatzker who acted as local research assistants in Hungary. Finally, we would like to express our gratitude to the research committee for the TOEFL® Young Students Series Research Grants (2016) and the staff at ETS, in particular, Veronika Timpe-Laughlin, for their support and prompt answers to our queries.

References

Adams, A. M., & Guillot, K. (2008). Working memory and writing in bilingual students. *International Journal of Applied Linguistics, 156*, 13–28.

Agresti, A. (2010). *Chapter 3 – Logistic regression models using cumulative logitsAnalysis of ordinal categorical data* (2nd ed.). New Jersey: Wiley.

Baddeley, A. D. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36*, 189–208.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *The Psychology of Learning and Motivation, 8*, 47–89.

Bailey, A. L. (2017). Theoretical and developmental issues to consider in the assessment of young learners' English language proficiency. In M. K. Wolf, & Y. G. Butler (Eds.). *English language proficiency assessments for young learners* (pp. 25–40). New York: Routledge.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models.* arXiv preprint arXiv,1506.04967.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Berninger, V. (1999). Coordinating transcription and text generation in working memory during composing: Automatic and constructive processes. *Learning Disability Quarterly, 22*, 19–112.

Bourdin, B., & Fayol, M. (2002). Even in adults, written production is still more costly than oral production. *International Journal of Psychology, 37*, 219–227.

Butler, Y. G. (2017). The role of affect in intraindividual variability in task performance for young learners. *TESOL Quarterly, 51*, 728–737.

Byrnes, H., & Manchón, R. M. (2014). Task, task performance, and writing development. In R. M. Manchón, & H. Byrnes (Eds.). *Task-based language learning – Insights from and for L2 writing* (pp. 267–299). Amsterdam: John Benjamins.

Cheng, L., Watanabe, Y., & Curtis, A. (Eds.). (2004). *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Lawrence Erlbaum.

Christensen, R. H. B. (2015). *Ordinal-regression models for ordinal data.* R package version 20156–28. https://cran.r-project.org/web/packages/ordinal/ordinal.pdf.

Conway, A. R. A., Kane, M. J., Bunting, M. F. D., Zach Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*, 769–786.

Cumming, A. (2016). Theoretical orientations to L2 writing. In R. Manchon, & P. K. Matsuda (Eds.). *Handbook of second and foreign language writing* (pp. 65–88). Berlin: Walter de Gruyter.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*, 5–43.

Cushing-Weigle, S. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition, 43*, 226–236.

Froetscher, D. (2016). A new national exam: A case of washback. In J. Banerjee, & D. Tsagari (Eds.). *Contemporary second language assessment* (pp. 61–81). London: Continuum.

Gathercole, S., & Alloway, T. P. (2008). *Working memory and learning: A practical guide for teachers.* London: Sage.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology, 40*, 177–192.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg, & E. R. Steinberg (Eds.). *Cognitive processes in writing* (pp. 3–30). Hillsdale, NJ: Erlbaum.

Hoskyn, M., & Swanson, H. L. (2003). The relationship between working memory and writing in younger and older adults. *Reading and Writing, 16*, 759–784.

Indrarathne, H. D. B. N., & Kormos, J. (2018). The role of working memory in processing L2 input: Insights from eye-tracking. *Bilingualism: Language and Cognition, 21*(2), 355–374.

Isbell, D. R. (2017). Assessing C2 writing ability on the Certificate of English Language Proficiency: Rater and examinee age effects. *Assessing Writing, 34*, 37–49.

Jackson, D. O., & Suethanapornkul, S. (2013). The Cognition Hypothesis: A synthesis and meta-analysis of research on second language task complexity. *Language Learning, 63*, 330–367.

Jarvis, H. L., & Gathercole, S. E. (2003). Verbal and non-verbal working memory and achievements on national curriculum tests at 11 and 14 years of age. *Educational and Child Psychology, 20*, 123–140.

Johnson, M. D. (2017). Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis. *Journal of Second Language Writing, 37*, 13–38.

Juffs, A., & Harrington, M. W. (2011). Aspects of working memory in L2. *Language Teaching, 44*, 137–166.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology General, 133*, 189–217.

Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy, & S. Ransdell (Eds.). *The science of writing: Theories, methods, individual differences and applications* (pp. 57–71). Mahwah, NJ: Lawrence Erlbaum.

Kellogg, R. T. (2008). Training writing skills: A cognitive development perspective. *Journal of Writing Research, 1*, 1–26.

Kim, Y., Payant, C., & Pearson, P. (2015). The intersection of task-based interaction, task complexity, and working memory. *Studies in Second Language Acquisition, 37*, 549–581.

Kormos, J. (2012). The role of individual differences in L2 writing. *Journal of Second Language Writing, 21*, 390–403.

Kormos, J. (2017). *The second language learning processes of students with specific learning difficulties.* New York: Routledge.

Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition, 11*, 261–271.

Kormos, J., Brunfaut, T., Michel, M (submitted). Motivational factors in computer-administered integrated skills tasks: A study of young learners.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review, 21*, 861–883.

Logan, G. D. (1994). On the ability to inhibit thought and action: A user's guide to the stop signal paradigm. In D. Dagenbach, & T. H. Carr (Eds.). *Inhibitory processes in attention, memory, and language* (pp. 189–239). San Diego, CA: Academic Press.

MacArthur, C. A., & Graham, S. (2016). Writing research from a cognitive perspective. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.). *Handbook of writing research* (pp. 24–40). New York/London: Guilford Press.

Manchón, R. M., Murphy, L., & Roca de Larios, J. (2009). Lexical retrieval processes and strategies in second language writing: A synthesis of empirical research. *International Journal of English Studies, 7*, 149–174.

McCutchen, D. (2011). From novice to expert: Implications of language skills and writing-relevant knowledge for memory during the development of writing skill. *Journal of Writing Research, 3*, 51–68.

Mitchell, A., Jarvis, S., O'Malley, M., & Konstantinova, I. (2015). Working memory measures and L2 proficiency. In Z. Wen, M. B. Mota, & A. McNeill (Eds.). *Working memory in second language acquisition and processing* (pp. 270–284). Bristol, UK: Multilingual Matters.

Nikolov, M. (Ed.). (2016). *Assessing young learners of English: Global and local perspectives.* Berlin: Springer.

O'Sullivan, B., & Weir, C. J. (2011). Test development and validation. In B. O'Sullivan (Ed.). *Language testing: Theories and practices* (pp. 13–32). Basingstoke: Palgrave Macmillan.

Oakhill, J., Hartt, J., & Samols, D. (2005). Levels of comprehension monitoring and working memory in good and poor comprehenders. *Reading and Writing, 18*, 657–686.

Olive, T. (2012). Working memory in writing. In V. Berninger (Ed.). *Past, present, and future contributions of cognitive writing research to cognitive psychology* (pp. 485–503). New York: Psychology Press.

Olive, T., Kellogg, R. T., & Piolat, A. (2008). Verbal, visual, and spatial working memory demands during text composition. *Applied Psycholinguistics, 29*, 669–687.

Papageorgiou, S., & Cho, Y. (2014). An investigation of the use of TOEFL® Junior™ Standard scores for ESL placement decisions in secondary education. *Language Testing, 31*, 223–239.

Papp, S., & Walczak, A. (2016). The development and validation of a computer-based test of English for young learners: Cambridge English young learners. In M. Nikolov (Ed.). *Assessing young learners of English: Global and local perspectives* (pp. 139–190). Berlin: Springer.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research.* Thousand Oaks CA: Sage.

R Core Team (2018). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Révész, A., Michel, M., & Lee, M.-J. (2017). *Investigating IELTS Academic writing Task 2 – Relationships between cognitive writing processes, text quality, and working memory*Retrieved from:. Australia: British Council, Cambridge English Language Assessment and IDP. https://www.ielts.org/teaching-and-research/research-reports/ielts_online_rr_2017-3.

Révész, A., Michel, M., & Lee, M.-J. (2019). Exploring second language writers' pausing and revision behaviors: A mixed methods study. *Studies in Second Language Acquisition* Special issue on methodological advances in investigating L2 writing processes.

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.). *Cognition and second language instruction* (pp. 287–318). Cambridge: Cambridge University Press.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics, 45*, 193–213.

So, Y., Wolf, M. K., Hauck, M. C., Mollaun, P., Rybinski, P., Tumposky, D., & Wang, L. (2015). *TOEFL Junior® Design Framework. ETS research report series*1–45 2015.

Steinlen, A. K. (2018). The development of German and English writing skills in a bilingual primary school in Germany. *Journal of Second Language Writing, 39*, 42–52.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate analysis.* London: Pearson.

Tannenbaum, R. J., & Baron, P. A. (2015). *Mapping scores from the TOEFL Junior® Comprehensive test onto the Common European Framework of Reference (CEFR). Research memorandum ETS RM–15-13.* Retrieved from: https://www.ets.org/Media/Research/pdf/RM-15-13.pdf.

Tu, Y. K., Kellett, M., Clerehugh, V., & Gilthorpe, M. S. (2005). Problems of correlations between explanatory variables in multiple regression analyses in the dental literature. *British Dental Journal, 199*, 457–461.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Oxford: Palgrave Macmillan.

Wolf, M. K., & Butler, Y. G. (2017). An overview of English language proficiency assessments for young learners. In M. K. Wolf, & Y. G. Butler (Eds.). *English language proficiency assessments for young learners* (pp. 3–22). New York: Routledge.

Woods, D. L., Kishiyama, M. M., Yund, E. W., Herron, T. J., Edwards, B., Poliva, O., Hink, R. F., & Reed, B. (2011). Improving digit span assessment of short-term verbal memory. *Journal of Clinical and Experimental Neuropsychology, 33*, 101–111.

Zalbidea, J. (2017). 'One task fits all'? The roles of task complexity, modality, and working memory capacity in L2 performance. *Modern Language Journal, 101*, 335–352.

Marije Michel is an assistant professor at Groningen University in the Netherlands and lecturer at Lancaster University. Her research focuses on socio-cognitive aspects of second language acquisition and task based language pedagogy. In her recent work she uses eye-tracking and key-stroke logging to investigate second language writing processes and alignment in written chat interactions.

Judit Kormos is a professor of Second Language Acquisition at the Department of Linguistics and English Language at Lancaster University. Her research interests include the psycholinguistic aspects of speech production, the role of attention and individual variables in language learning and special educational needs in foreign language education.

Tineke Brunfaut is a senior lecturer in the Department of Linguistics and English Language at Lancaster University. Her main research interests are in language testing, and reading and listening in a second/foreign language. She is a recipient of the *ILTA Best Article Award,* the *e-Assessment Best Research award,* and the *TOEFL Outstanding Young Scholar Award.*

Michael Ratajczak is a specialist in multilevel modelling. He completed an MSc in Psychological Research Methods and a BSc in Psychology in Education at Lancaster University. Currently, he is a PhD student in the Department of Linguistics and English Language at Lancaster University, funded by the UK's Economic and Social Research Council (ESRC) and by the UK's National Health Service (NHS).